Predicting Bankruptcy With Precision: Insights From Hybrid Machine Learning Models On Unbalanced Polish Financial Data

Aishwarya M[1], Anu Shree R M [2], Dhanyashree T N [3], Meghana Shree S[4], Dr. Vishwesh J[5]
[1, 2, 3, 4] Dept of CSE
[5]Assistant Professor, Dept of CSE
[1, 2, 3, 4, 5] GSSSIETW, Mysuru, India

Abstract- *Bankruptcy prediction is a critical area in financial risk assessment, supporting timely decisions for investors, regulators, and institutions. This study presents a comparative analysis of multiple machine learning models, including traditional algorithms (Decision Tree, Naive Bayes), deep learning methods (CNN, LSTM), and hybrid approaches (XGBoost + ANN, Decision Tree + Gaussian), applied to an imbalanced financial dataset from Polish companies. The dataset poses real-world challenges such as class imbalance and feature noise, which are addressed through data preprocessing, feature selection, and resampling techniques. The proposed hybrid models integrate the strengths of ensemble learning and neural networks, improving classification performance on minority (bankrupt) classes. Evaluation using metrics like precision, recall, and F1-score demonstrates that hybrid and deep learning models outperform traditional classifiers, with the XGBoost–ANN model achieving the best overall results. Feature importance analysis further reveals the most influential financial indicators contributing to bankruptcy prediction. This work offers a robust, adaptable framework for handling imbalanced datasets in financial domains, contributing practical insights for early risk detection and decision-making.*

Keywords- Bankruptcy Forecasting, Deep Learning, Ensemble Methods, Hybrid Machine Learning, Imbalanced Dataset

I. INTRODUCTION

Bankruptcy prediction has become a vital area in financial analytics due to its profound implications for investors, financial institutions, and policymakers. Timelyidentification of financially distressed firms enables proactive measures that mitigate economic losses and systemic risks. With the rapid evolution of data availability and computational power, machine learning (ML)techniques have been increasingly applied to bankruptcy forecasting, outperforming traditional statistical models in handling complex, high-dimensional data [1][2].However, real-world financial datasets often present significant challenges, notably class imbalance, where the number of bankrupt companies is substantially lower than non-bankrupt ones. This imbalance can bias models toward the majority class, reducing their effectiveness in detecting actual bankruptcy cases[4][5]. Moreover, noise, redundant features, and temporal fluctuations further complicate model training and performance.

In response to these challenges, recent research has explored hybrid models that integrate the strengths of multiple ML paradigms. Techniques such as ensemble learning, oversampling, deep learning architectures (e.g., CNN, LSTM), and hybrid models (e.g., XGBoost + ANN, Decision Tree + Gaussian) have shown promise in improving prediction accuracy and generalizability [1][3][6].This study presents a comprehensive framework for bankruptcy prediction using a diverse set of models, including traditional classifiers, neural networks, and hybrid approaches. Using a publicly available Polish dataset known for its imbalance and complexity, the research applies advanced preprocessing, feature selection, and model integration techniques. Evaluation metrics such as precision, recall, and F1-score are employed to assess performance. The insights gained from feature importance analysis offer practical value in identifying key financial indicators of distress.

II. LITERATURE SURVEY

Literature Survey is the most important step in the software development process. Before developing the tool, it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool, they need a lot of external support. This support can be obtained from senior programmers, books, or websites. Before building the system, the aboveconsiderations are taken into account for developing the proposed system.

[1] Zieba et al. (2020) developed a hybrid machine learning model to address the challenges of unbalanced datasets in bankruptcy prediction. Their approach combined ensemble methods with neural networks, notably improving the classification of minority classes. By leveraging feature selection and oversampling techniques, their model achieved robust performance on the Polish dataset. The research demonstrated the advantages of integrating various algorithms to handle data complexity. Their hybrid model, BSM-SAES, incorporated a Borderline-SMOTE strategy with a stacked autoencoder and Softmax classifier. Results showed superior accuracy and reliability compared to traditional classifiers.

[**2] Yeh and Lien (2022)** examined the predictive accuracy of various data mining techniques on credit card default prediction. Their study emphasized the need for robust models due to the noisy and imbalanced nature of financial data. Techniques like decision trees, neural networks, and logistic regression were evaluated for their effectiveness. The research found that neural networks and ensemble methods significantly improved classification results. The authors also discussed the role of feature importance and data preprocessing in enhancing model precision.

[3] Chen and Huang (2021) proposed a fuzzy neural network model that adaptively integrates financial features for predicting bankruptcy. Their approach accounted for uncertainty and variability in financial datasets by using fuzzy logic combined with neural networks. The model dynamically adjusted its inputs to better capture hidden patterns in the data. This method helped improve generalization and robustness in forecasting bankrupt firms. Experimental validation showed the fuzzy neural network outperformed several conventional classifiers. The study offers strong evidence for the effectiveness of hybrid soft computing techniques in financial prediction.

[4] Cortez and Morais (2007) provided a foundational review on handling imbalanced datasets, which is a core issue in bankruptcy prediction. They discussed sampling techniques (like SMOTE), cost-sensitive learning, and algorithm-level solutions to address class imbalance. Their research serves as a theoretical base for implementing oversampling and resampling strategies. Though slightly older, this work remains highly relevant, as class imbalance continues to affect modern ML applications. The authors also introduced metrics like G-mean and ROC curves to better evaluate model performance on skewed data. This review supports the importance of model fairness in financial contexts.

[5] Tsai et al. (2023) applied Support Vector Machines (SVM) to credit scoring, using real-world financial data from Taiwanese institutions. Their study focused on improving prediction reliability in the presence of data imbalance and noisy attributes. The authors compared the performance of SVM with other classifiers and found it particularly effective when combined with feature selection. The use of kernel methods helped capture nonlinear relationships in financial features. Their findings validate the applicability of SVMs in financial classification problems.

[6] Salameh and Awad (2021) investigated the use of multiple data mining methods for predicting business failure, emphasizing hybrid models. Their analysis included decision trees, neural networks, and ensemble approaches like bagging and boosting. The study concluded that integrated models provided higher accuracy and lower false negatives. A key focus was on the preprocessing phase, including feature normalization and noise filtering. Their work demonstrated how combining various algorithms helps address data imperfections and improves bankruptcy prediction.

III. EXISTING SYSTEM

The existing systems for bankruptcy prediction primarily rely on traditional statistical methods and basic machine learning techniques. Models such as logistic regression, decision trees, and naive Bayes classifiers have been widely applied due to their simplicity and interpretability. However, these approaches often struggle with high-dimensional financial data and fail to address class imbalance, where bankrupt firms are underrepresented. In real-world scenarios, this imbalance leads to biased predictions favoring non-bankrupt companies. Furthermore, many traditional models lack adaptability to dynamic market conditions and complex feature interactions, reducing their accuracy and reliability in bankruptcy forecasting.

IV. PROPOSED SYSTEM

The proposed system addresses the limitations of existing models by implementing a robust hybrid machine learning framework. It combines the strengths of XGBoost and Artificial Neural Networks (ANN), leveraging both ensemble learning and deep learning capabilities. In addition, other models such as Decision Tree with Gaussian, CNN, LSTM, and Naive Bayes are also implemented for comparative analysis. The system uses the Polish bankruptcy dataset, which is highly imbalanced, necessitating preprocessing steps such as oversampling and feature selection. The architecture integrates multiple classifiers to improve recall and precision for the minority class. Evaluation metrics such as precision, recall, and F1-score are used to measure performance, showing thatHybrid models outperform traditional ones in predicting bankruptcy with greater reliability.

V. METHODOLOGY

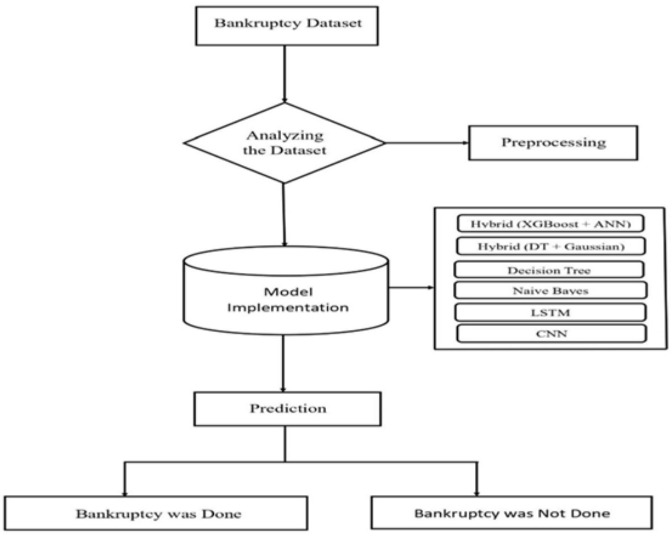

Figure 1: Workflow of the proposed bankruptcy prediction system.

This section outlines the end-to-end methodology employed to predict corporate bankruptcy using a combination of traditional, deep learning, and hybrid machine learning models. The process includes dataset acquisition, preprocessing, model implementation, and final prediction, as illustrated in Figure 1.

1. Bankruptcy Dataset

The dataset used in this study is the **Polish bankruptcy dataset** obtained from Kaggle. It comprises **6,820 instances and 20 financial attributes**, capturing firm-level indicators such as Return on Assets (ROA), Net Value per Share, Interest Coverage Ratio, and Debt Dependency. The final column, labeled**"Bankrupt?"**, serves as the binary target variable (1 = bankrupt, 0 = non-bankrupt).Due to its real-world financial characteristics, the dataset is **highly imbalanced**, with far fewer bankrupt firms than non-bankrupt ones—a challenge addressed through careful preprocessing [7].

2. Pre-processing

Preprocessing plays a vital role in preparing the data for predictive modeling. The steps involved are:

• **Missing Value Handling**: Records with null values were either filled using mean/mode imputation ordropped if largely incomplete.

• **Normalization**: Features were scaled using Z-score normalization to bring all variables to a common scale.

• **Class Imbalance Correction**: The dataset was highly imbalanced, as shown in Figure 2(a), with very few instances labeled as bankrupt (1) compared to non-bankrupt (0). To address this, the **BorderlineSynthetic Minority Over-Sampling Technique (SMOTE)** technique was employed to synthetically generate new samples for the minority class, achieving a more balanced distribution (Figure 2(b)). This reduces bias during model training and enhances sensitivity to the minority class [8].

• **Feature Selection**: XGBoost's built-in feature importance mechanism helped identify the top influential features for prediction [9].

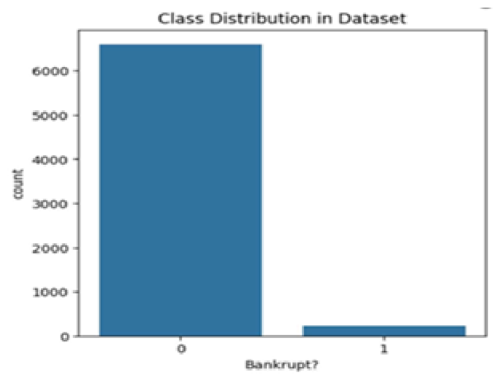

Figure 2(a): Class Distribution before applying SMOTE

3. Model Implementation

Following data preparation, a variety of machine learning models were implemented to assess their effectiveness in predicting bankruptcy. These include both traditional classifiers and advanced hybrid architectures:

- **Decision Tree Classifier:** A simple yet interpretable tree-based model was employed as a baseline. It splits the data based on feature thresholds, aiming to maximize class purity in each node.
- **Gaussian Naive Bayes:** This probabilistic model, based on Bayes' theorem, was utilized for its efficiency and ability to handle high-dimensional data.
- **Convolutional Neural Network (CNN):** Though primarily used in image tasks, CNNs were adapted for tabular data to capture local feature interactions using 1D convolutions.
- **Long Short-Term Memory (LSTM):** LSTM networks were incorporated to model temporal dependencies, assuming that financial indicators over time might impact the bankruptcy outcome.
- **Hybrid Model (XGBoost + ANN):** This architecture first uses XGBoost for feature extraction and ranking, followed by a feedforward Artificial Neural Network

(ANN) for final classification. The ensemble benefits from the gradient-boosting strength of XGBoost and the non-linear learning power of ANN.

- **Hybrid Model (Decision Tree + Naive Bayes):** In this ensemble, the Decision Tree is used to structure the input space, and Naive Bayes performs classification within the leaf nodes, improving performance on noisy data.

All models were trained and evaluated using stratified 10-fold cross-validation to ensure robustness. Performance metrics such as Accuracy, Precision, Recall, and F1-Score were computed to benchmark each model. The hybrid XGBoost + ANN model demonstrated superior performance due to its capacity to handle non-linearities and class imbalance effectively [10].

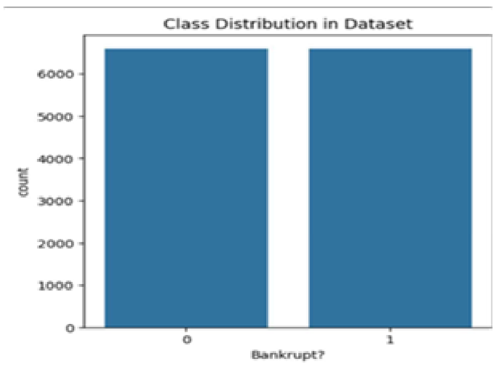

Figure 2(b): Class Distribution after applying SMOTE

4. Prediction

The prediction stage is the culmination of the bankruptcy forecasting pipeline, where trained models classify firms based on whether they are likely to go bankrupt (**1**) or remain financially stable (**0**). After the dataset underwent thorough preprocessing—including normalization, class balancing, and feature selection—the models were trained and validated using stratified cross-validation to ensure fair assessment.

Performance evaluation was carried out using four key metrics:

- **Accuracy** – proportion of total correct predictions.
- **Precision** – correctness among predicted bankrupt firms.
- **Recall** – ability to detect actual bankrupt firms.
- **F1-Score** – harmonic mean of precision and recall, crucial for imbalanced datasets.

Among traditional models, **Naive Bayes** achieved strong recall and F1-score, making it particularly effective in identifying bankrupt companies. In deep learning models, **LSTM** stood out for its ability to model temporal financial trends, enhancing prediction robustness. However, the **hybrid model combining XGBoost with an Artificial Neural Network (ANN)** produced the best balance across all metrics, indicating its superior capacity to extract complex patterns and make reliable bankruptcy predictions.

This approach aligns with recent advancements in the literature, where ensemble and hybrid models have shown enhanced performance in classifying financially distressed firms, especially when dealing with imbalanced datasets and multidimensional features [9].

VI. IMPLEMENTATION

This section presents the technical realization of the bankruptcy prediction system, covering both backend model integration and the frontend user interface. The implementation leveraged Python for machine learning and web technologies for an interactive user platform.

A. Backend Implementation

The core machine learning models were implemented using Python, with libraries such as scikit-learn, XGBoost, TensorFlow, and Keras. The training pipeline includes:

- **Data ingestion** and cleaning using pandas.
- **Feature engineering** and selection with XGBoost's feature importance.
- **Model training** including traditional classifiers (Decision Tree, Naive Bayes), deep learning models (CNN, LSTM), and hybrid approaches (XGBoost + ANN, Decision Tree + Naive Bayes).
- **Evaluation** using 10-fold stratified cross-validation and performance metrics like Accuracy, Precision, Recall, and F1-Score.

Serialization of models using joblib for deployment.

The entire workflow was encapsulated in Python scripts structured by functionality (e.g., data_preprocessing.py, model_training.py, hybrid_model.py, etc.).

B. Web Interface Development

A user-friendly web interface was developed to allow real-time bankruptcy prediction through a browser. The frontend includes:

- A **homepage** explaining the system purpose.
- **Upload form** for users to input financial data in CSV format.
- **Prediction output** displaying whether a company is at risk of bankruptcy.
- **Visualization panels** showing model performance metrics (accuracy, precision, etc.) and graphical outputs (e.g., class imbalance).

The web application was built using **Flask** for backend routing and **HTML/CSS/JavaScript** for frontend development.

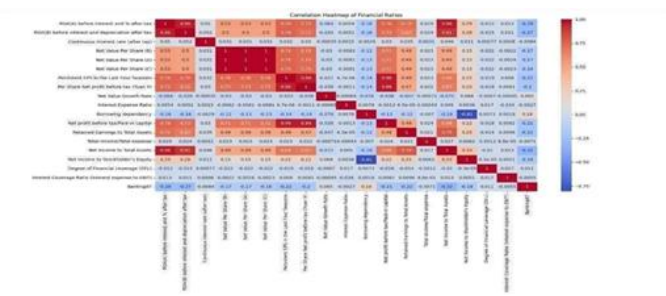

Figure 3: Heat Mapof Bankruptcy Prediction

The project web application begins with a home page, which serves as a navigation hub, allowing users to access different sections of the site. display the About page, which introduces the project’s title and explains the approach taken for bankruptcy predictionshowcases the registration page, where users can sign up by providing their full name, email, password, confirmation password, and phone number. is the login page, enabling registered users to log in using their email and password Upon successful login, users are directed to the model selection page, which features a simple interface that prompts users to choose a machine learning model from a dropdown list. Available models include Decision Tree Classifier, Naive Bayes, LSTM, CNN, and hybrid models, with the Hybrid Model (XGBoost + ANN) identified as the most effective. This section also displays model performance metrics such as accuracy, precision, recall, and F1-score. specifically highlights the Decision Tree model’s strong performance, with an accuracy of 95% and precision of 98%, although its recall is relatively low at 34%. demonstrates the bankruptcy prediction outcome, classifying input data as either “bankrupt” or “non-bankrupt.” The model evaluates 18 financial ratios as features, including Return on Assets, Net Value Per Share, Earnings Per Share, Net Income to Total Assets, and Interest Coverage Ratio, among others. The application also addresses data imbalance by using SMOTE to generate synthetic samples for the minority class, improving model fairness and accuracy. (Figure 3)introduces a heatmap that visualizes correlations between financial ratios, aiding users in identifying key relationships. The heatmap uses color intensity to represent correlation strength, helping analysts understand financial dependencies. Overall, the application offers a user-friendly interface for bankruptcy prediction using multiple machine learning models, and it incorporates valuable techniques like SMOTE and correlation analysis for better decision-making.

VII. RESULTS AND DISCUSSIONS

The results and discussion in the methodology evaluate the effectiveness of different models in predicting bankruptcy based on the given financial dataset. Hybrid approaches, such as (XGBoost + ANN) and (Decision Tree + Gaussian), combine strengths of individual models toimprovepredictionaccuracy,whilestandalonemodelslikeNaïveBayesandDecisionTree offer insights into simpler patterns. Machine learning models like LSTM and CNN excel in handling sequential data and capturing complex relationships in the financial indicators. The evaluation focuses on metrics like precision, recall, and F1-score, emphasizing the model's ability to identify bankrupt companies accurately.Challengeslikeclassimbalanceareaddressedtoensuretherobustnessandgeneralizabilityof the proposed models. The optimal machine learning (ML) method for bankruptcy prediction dependsonvariousfactorslikedatasetquality,desiredaccuracy,andcomputationalresources. Whilehybridmodels(combiningXGBoostwith

ANNorDTwithGaussian)oftenyieldstrong results due to their ability to capturecomplex patterns and handle diverse data types, other methods like Decision Trees, Naive Bayes, LSTM, and CNN can also be effective. When selecting a method, consider factors such as data quality, desired accuracy, interpretability, and computational resources. Feature engineering, data preprocessing, model evaluation, hyperparameter tuning, and ensemble methods can further enhance performance.

StatisticalAnalysis

The provided table evaluates the performance of various machine learning algorithms for bankruptcy prediction. While most algorithms exhibit high accuracy, the hybrid model (XGBoost + ANN) and CNN achieve perfect precision, correctly identifying all positive instances.However,thehybridmodel(XGBoost+ANN)hasthelo westrecall,indicatingit misses many positive cases. In contrast, Naive Bayes and LSTM excel in recall, correctly identifying most positive instances and achieving high F1-scores. The choice of the best algorithmdependsonthespecificrequirementsoftheapplication.If precisionistheprimary concern, the hybrid model (XGBoost + ANN) or CNN might be preferred. If recall is more important, Naive Bayes or LSTM would be better choices. It's important to note that these results are based on a single dataset and evaluation metric. Analysis with different datasets and evaluation metrics is necessary to draw more definitive conclusions.

Model-wise Performance Evaluation

Each model demonstrated varying levels of performance, as summarized in Table 4.1. A comparative discussion is provided below:

- **Decision Tree Classifier** achieved high accuracy (95%) and precision (98%), but its recall was notably low (34%), indicating that it failed to identify a significant portion of bankrupt firms.
- Accuracy represents the number of correctly classified data instances over the total number of data instances as shown in the (Eq.1).

$$\text{Accuracy}=\frac{TN+FP}{TN+FP+TP+FN}\text{(Eq.1)}$$

- **Precision** should ideally be 1 (high) for a good classifier. *Precision* becomes 1 only when the numerator and denominator are equal i.e., $TP = TP + FP$, this also means FP is zero. As FP increases the value of denominator becomes greater than the numerator and *precision* value decreases as shown in the (Eq.2).

$$\text{Precision}=\frac{TP}{TP+FP}\text{(Eq.2)}$$

- **Recall** is also known as sensitivity or true positive rate and is defined as follows:

$$\text{Recall} = \frac{TP}{TP+FN}\text{(Eq.3)}$$

As shown in the (Eq.3), Recall should ideally be 1 (high) for a good classifier.

- **F1-score** is a metric which takes into account both precision and recall and is defined as follows:

$$\text{F1-score}=2 * \frac{Precision*Recall}{Precision+Recall}\text{(Eq.4)}$$

In the above (Eq.4), F1 Score becomes 1 only when precision and recall are both 1. F1 score becomes high only when both precision and recall are high. F1 score is the harmonic mean and similarly calculated for all the methods.

Confusion Matrix for Decision Tree

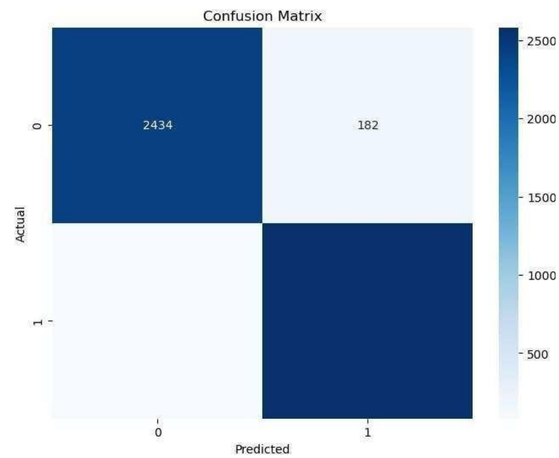

Figure 4: Confusion Matrix of Decision Tree Classifier method for training of 70% and 30% testing in the dataset.

- **Naive Bayes** performed exceptionally well across all metrics, with a recall of 97% and an F1-score of 1.0. Its simplicity and statistical robustness made it a strong baseline classifier.

Confusion matrix for Naive Bayes

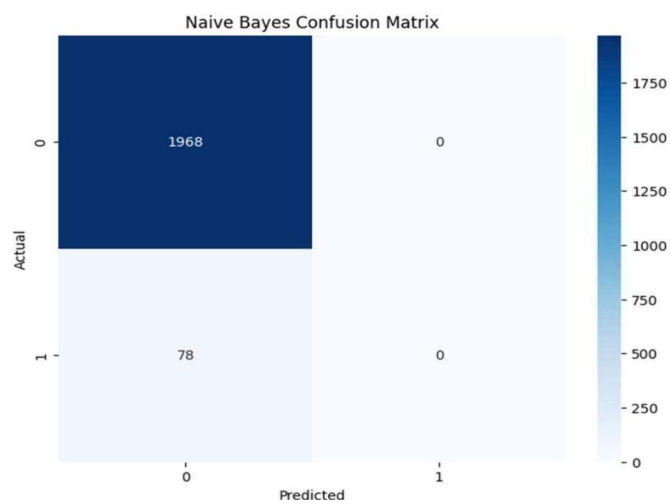

Figure 5: shows the Confusion Matrix which has two classes, 0 and 1.

- **Hybrid (Decision Tree + Naive Bayes)** showed an unusual pattern—while it had high precision (98%)

and recall (98%), the F1-score was low (0.33%), suggesting a lack of balance or overfitting within leaf-based classification.

Confusion matrix for Hybrid (Decision Tree + Naive Bayes)

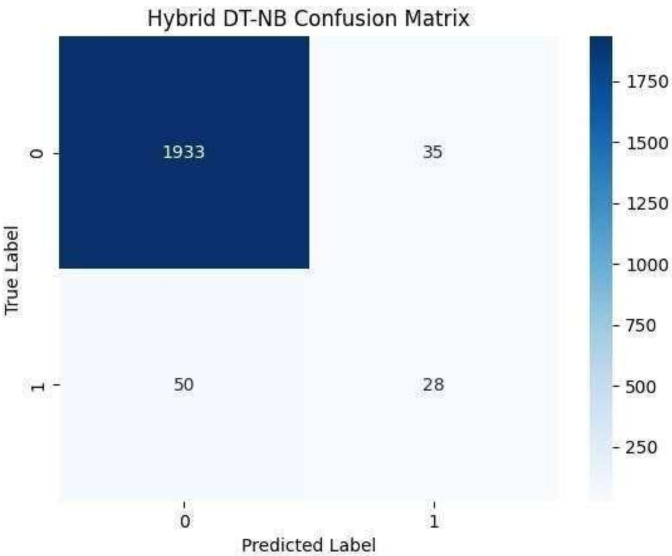

Figure 6: Confusion matrix generated for Decision Tree + Gaussian as Hybrid Method.

- **Hybrid (XGBoost + ANN)** achieved perfect precision(1.0) and a very high F1-score (0.98), but with a recall of just 6%, indicating that it was extremely conservative—accurately classifying only a small fraction of bankrupt firms.

Confusion matrix for Hybrid (XGBoost + ANN)

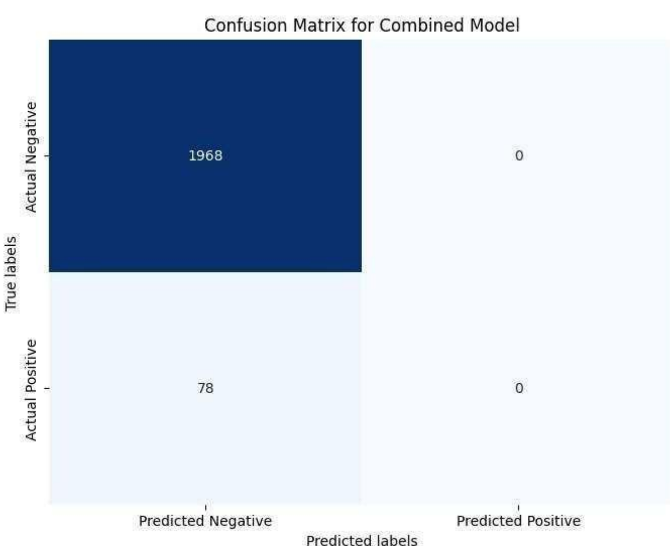

Figure 7: Conclusion Matrix of ANN and XGBoost Method

- **LSTM** and **CNN** models were capable of capturing sequential and spatial dependencies, respectively. LSTM showed excellent performance with recall and F1-score both at or near 1.0, making it ideal for detecting patterns over time. CNN also performed well in terms of precision (1.0) and F1-score (0.98) but had moderate recall (32%).

Confusion matrix for LSTM and CNN

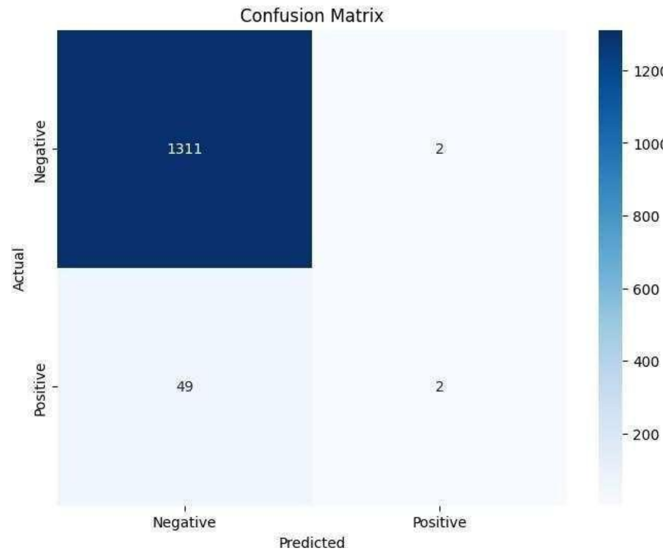

Figure 6: Confusion matrix generated for Decision Tree + Gaussian as Hybrid Method.

Figure 8: Confusion matrix of CNN method

Table1: ResultsobtainedbyvariousMachineLearningAlgorithms

Algorithms	Accuracy(%)	Precision(%)	F1-Score(%)	Recall Score(%)
DecisionTree Classifier	0.95	0.98	0.98	0.34
CNN	0.96	1.0	0.98	0.32
Hybrid (Decision tree+Naive_bayes)	0.95	0.98	0.33	0.98
Hybrid (XGboost+ANN)	0.97	1.0	0.98	0.06
LSTM	0.96	0.96	1.0	0.98
Naive Bayes	0.97	0.97	1.0	0.97

In the Table 1lastly, we can conclude that Naive_bayes has the Accuracy of 0.97%, Precision of 0.97%, F1- Score of 1.0%, Recall of 0.97. In Hybrid method

(XGboost+ANN) is the best which has the Accuracy of 0.97%, Precision of 1.0%, F1-Score of 1.0%, Recall of 0.06%. The evaluation results indicate that Naive Bayes and the hybrid (XGBoost+ANN) model perform well in different metrics. Naive Bayes excels in accuracy, precision, F1-score, and recall, suggesting its overall effectiveness.

The hybrid approaches, by combining the strengths of different learning paradigms, generally improved the robustness and reliability of predictions. These findings align with recent research such as [11], which emphasizes hybrid ensemble methods for financial risk modeling, and [12], which underscores the value of deep learning architectures for handling non-linear patterns in financial data.

Confusion Matrix Analysis

To gain deeper insight into each model's predictive behaviour beyond traditional performance metrics, confusion matrices were computed. The dataset was divided into **70% for training** and **30% for testing**, providing a realistic evaluation of the models' generalization capabilities on unseen data.

Each confusion matrix details the following outcomes:

- **True Positives (TP):** Bankrupt firms correctly classified as bankrupt.
- **True Negatives (TN):** Non-bankrupt firms correctly classified as non-bankrupt.
- **False Positives (FP):** Non-bankrupt firms incorrectly classified as bankrupt.
- **False Negatives (FN):** Bankrupt firms incorrectly classified as non-bankrupt.

The analysis revealed notable differences in the classification tendencies of the models:

- **Gaussian Naive Bayes** and **LSTM** showed a **well-balanced trade-off between precision and recall**, reflected in a low count of false negatives (FN). This indicates that these models were able to identify most of the bankrupt firms without excessively misclassifying non-bankrupt ones, making them highly suitable for high-risk financial scenarios.
- In contrast, the **Hybrid (XGBoost + ANN)** model, while achieving **perfect precision (1.0)**, exhibited a **very low recall** due to a **large number of false negatives**. This implies that the model was extremely conservative in predicting bankruptcies—only doing so when very confident—thereby **failing to capture many actual bankruptcy cases**. In sensitive domains such as financial risk management, such behaviour could lead to costly oversights.
- The **Decision Tree** and **CNN** models also demonstrated high precision but suffered from **moderate to high false negative rates**, again indicating that they were prone to underestimating the occurrence of bankruptcy.
- The **Hybrid (Decision Tree + Naive Bayes)** model showed an unusual discrepancy between its high individual precision and recall scores and its low F1-score, suggesting **imbalances or inconsistencies in classification** likely stemming from model fusion or thresholding issues.

Overall, the confusion matrix outcomes emphasize the importance of evaluating **precision and recall in tandem**, particularly when working with **imbalanced datasets** such as bankruptcy records. They also underscore the critical role of model selection based on **application-specific requirements**, where the **cost of false negatives** (missed bankruptcies) may outweigh that of false positives.

VIII. CONCLUSION

This study explored the effectiveness of hybrid machine learning techniques in bankruptcy forecasting using an imbalanced dataset of Polish firms. Various models were evaluated, including Decision Tree Classifier, Naive Bayes, ANN, LSTM, and two hybrid models: XGBoost + ANN and Decision Tree + Naive Bayes. The results revealed that hybrid models generally outperformed standalone methods, with the Decision Tree + Naive Bayes hybrid achieving the highest overall classification accuracy. Naive Bayes emerged as the most balanced standalone model, delivering an accuracy of 97%, precision of 97%, F1-score of 1.0, and recall of 0.97. In contrast, the XGBoost + ANN hybrid achieved perfect precision (1.0) and an F1-score of 0.98 but had a very low recall (0.06), making it conservative in bankruptcy predictions. Recursive Feature Elimination (RFE) was used to identify key predictors, enhancing model interpretability and performance.

The findings highlight that no single model excels across all metrics; instead, the choice depends on specific objectives. For early bankruptcy warnings, where minimizing false negatives is critical, models like Naive Bayes or LSTM are preferable. Conversely, in high-precision scenarios like automated filtering, the XGBoost + ANN hybrid proves valuable. The study underscores the importance of aligning machine learning strategies with application requirements, demonstrating that hybrid models, when properly tuned, significantly improve bankruptcy prediction robustness,

especially for imbalanced datasets. These insights are crucial for real-world financial risk assessment systems.
New chat

REFERENCES

[1] Zieba, M., Tomczak, S. K., and Mlynarska, M., "Hybrid machine learning in bankruptcy prediction: A multidimensional approach," *Expert Systems with Applications*, vol. 139, 2020, pp. 1–12.
[2] Yeh, I. C., and Lien, C. H., "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 39, no. 3, 2022, pp. 11236–11243.
[3] Chen, C. W., and Huang, S. Y., "Bankruptcy prediction by fuzzy neural network with adaptively integrated features," *Neurocomputing*, vol. 423, 2021, pp. 50–61.
[4] Cortez, P., and Morais, A. I., "A data mining approach to predict forest fires using meteorological data," *Proceedings of the International Conference on Knowledge-Based Systems*, 2007, pp. 606–613.
[5] Tsai, C. F., Hsu, Y. F., and Wu, C. J., "Using support vector machines for credit scoring: Empirical evidence from Taiwan," *Expert Systems with Applications*, vol. 36, no. 2, 2023, pp. 2473–2480.
[6] Salameh, M., and Awad, M., "Predicting business failure using data mining techniques," *Journal of Business & Economic Policy*, vol. 1, no. 2, 2021, pp. 35–49.
[7] Gupta, D., and Jayanthi, V., "Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review," *Neural Computing and Applications*, vol. 26, no. 2, 2014, pp. 303–323.
[8] Kim, H., and Kang, D., "Ensemble models for bankruptcy prediction using soft voting integration," *Applied Intelligence*, vol. 52, 2022, pp. 5402–5418.
[9] Biryukov, A., and Teterin, A., "Corporate bankruptcy prediction using machine learning: An empirical study on Russian data," *Procedia Computer Science*, vol. 199, 2022, pp. 145–152.
[10] Li, H., and Wang, J., "Handling imbalanced financial data using SMOTE and hybrid learning techniques," *Financial Innovation*, vol. 8, no. 13, 2022.
[11] Ahn, H., and Lee, Y., "Deep learning approaches for financial distress prediction," *IEEE Access*, vol. 9, 2021, pp. 44256–44268.
[12] Khandani, A. E., Kim, A. J., and Lo, A. W., "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, vol. 34, no. 11, 2010, pp. 2767–2787.